ELSEVIER

Theoretical Computer Science 201 (1998) 151–170

Theoretical Computer Science

Inventories of unavoidable languages and the word-extension conjecture

Laurent Rosaz *

L.I.T.P. (Laboratoire d'Informatique Théorique et de Programmation), Université Paris VII, 2 place Jussieu, 75251 Paris Cedex 05, France

Received October 1992; revised November 1996
Communicated by D. Perrin

Abstract

A language X on an alphabet A is unavoidable iff all but finitely many words in A^* have a factor in X. In this paper, I prove that the inventory of unavoidable languages of n words can be explicitly made for every n, that the reduced unavoidable languages of given cardinality are finite in number (an unavoidable language is minimal if no proper subset is unavoidable, it is reduced if it is minimal and if whenever a word is replaced by a proper factor, the resulting unavoidable language is not minimal), and I give a counterexample to the word-extension conjecture (which said that in every unavoidable language, there is a word w and a letter a, such that the language, where w is replaced by wa, is still unavoidable).

1. Introduction

A language X on the finite alphabet A (that is a subset X of the free monoid A^*, that is a set X of words on an alphabet A) is unavoidable iff all but finitely many words in A^* have a factor in X. This is not to be confused with unavoidable patterns, such as the square in square-free words. See [2, 15] for references on this latter topic.

Unavoidable languages appeared in 1964 in a paper by Schutzenberger [21] where he gave a bound on the maximal length of a word that avoids a finite unavoidable language. This bound depends on the maximal length of the words in the unavoidable language. Crochemore et al. proved later (in 1983) in [5], that the bound given by Schutzenberger was the best possible.

Unavoidable languages were explicitly introduced in 1983 by Ehrenfeucht et al. in [7] in a generalization of Higman's result, [9]. Higman's theorem states that if A is a finite alphabet, then, in every infinite language $\{u_i \mid i \in I\}$ on A, there is a pair (u_i, u_j) of words with $i \neq j$ such that u_i is a sub-word of u_j (a *sub-word* of a word u is a word

* Present address: LRI, bat. 490, Univ. Paris Sud, 91405 Orsay. E-mail: rosaz@lri.fr.

0304-3975/98/$19.00
PII S0304-3975(97)00031-5

v obtained by taking a subsequence of the letters of u. For example ac is a sub-word of abc). The generalization by Ehrenfeucht et al. says that if the partial order relation $\leqslant$ is the reflexive transitive closure of:

"$u \leqslant v$ iff $\exists w, y, z$ with $w, z \in A^*$ and $y \in X$ such that $u = wz$ and $v = wyz$",

then X is unavoidable iff every infinite language on A contains two different words u and v such that $u \leqslant v$. One gets Higman's theorem from this result by considering $X = A$.

Bucher et al. generalized the latter result in [3]. Kruskal in [10] and Puel in [18] gave some similar results on trees instead of words.

It had been conjectured that if X is unavoidable, then there is a word w in X and a letter α such that $X - \{w\} + \{w\alpha\}$ is still unavoidable. This word-extension conjecture was often called Ehrenfeucht's conjecture, though it might be due to Haussler. A counter-example to this conjecture will be given in this paper.

In 1984, Choffrut and Culik published [4] where they recalled some basic results (an unavoidable language always contains a finite sublanguage which is unavoidable, recall of the automaton of Aho and Corasick [1] and use of this automaton to decide whether a given language is unavoidable) and gave some interesting new ones (partial answer to the word-extension conjecture, uniqueness of the extension of a word when it exists, first use of some important tools such as bi-finite periodic words...). This paper is the one to be read as an introduction to unavoidable languages.

In [19], I introduced cuts which are closely related to unavoidable languages.

There are other notions of "unavoidable" in theoretical computer science: Unavoidable patterns such as the square in square-free words, see [2] or [15] for references on this topic; unavoidable words with patterns, see [13]; unavoidable trees, which were studied by Puel in [18] where she generalized Kruskal's theorem [10]; unavoidable subset of an ordered set, see [17].

This paper deals with inventories of unavoidable languages of fixed cardinality and with the word-extension conjecture. In order to test systematically the word-extension conjecture, one needs an efficient way to make inventories of unavoidable languages. I will prove that unavoidable languages of fixed cardinality can be described in a sense I will define. The description can be computed using a recursive algorithm, so that the explicit description of the unavoidable languages of given cardinality can be made. The shape of the descriptions enables testing the word-extension conjecture for given cardinality, so that one can look systematically for a counter-example. Some remarks allow significant simplification in the explicit description (which is somewhat bulky). These remarks can be found in [20]. The computation leads to a counter-example of seven words which will be given. As a corollary of the inventories, I will answer Choffrut's question and prove that reduced unavoidable languages of given cardinality are finite in number. (An unavoidable language X is minimal iff no proper sublanguage Y of X is unavoidable. A minimal unavoidable language is reduced iff for every word $x \in X$ and for every factor y of x different from x, the unavoidable language $X - \{x\} + \{y\}$ is not minimal.)

I will first recall in Section 2 some basic definitions on words: on finite words and languages (length of a word, word ε, concatenation, factors, prefixes, suffixes on finite words, operations $+$, product, * and $^+$ on languages), on infinite words (finite factor of a bi-infinite word, periodic bi-infinite words, notation $u^{\mathbb{Z}}$, equivalence $\equiv$ (equality up to a translation)) and on automata (finite or bi-finite word recognized by an automaton, by a path in the automaton, notation $\xrightarrow{u}{}^*$). I will define unavoidable languages in Section 3.1, I give some examples in Section 3.2 and I describe the unavoidable languages of less than three words (on the alphabet $A=\{a,b\}$) in Section 3.3. In Section 4, I will introduce descriptions and claim that the set of unavoidable languages of less than N words is describable for every integer N. I introduce the notation $u^\bullet$ and I define descriptions as well as languages described by a description. I claim and prove that unavoidable languages are describable. In Section 5, I will show that reduced unavoidable languages of fixed cardinality are finite in number. The definitions of minimal and of reduced unavoidable languages are given in Section 5.2, then in Section 5.3, I will claim and prove that the reduced unavoidable languages are finite in number. The proof uses a lemma (known as Dickson's lemma) given in Section 5.1. Section 6 deals with the word-extension conjecture. Section 6.1 gives as a corollary of the theorem of Section 4, that this conjecture is decidable for unavoidable languages of fixed cardinality. Then I will give in Section 6.2 a counter-example that one can find by making the list of unavoidable languages of seven words. In Section 7, I will give a few open problems.

2. Basic definitions

To begin with, let me be precise that I consider that $\mathbb{N}=\{0,1,\ldots\}$, so that $0\in\mathbb{N}$ (the set $\{1,2,\ldots\}=\mathbb{N}-\{0\}$ will be denoted by $\mathbb{N}^*$). I also specify that whenever I write $X-Y$, where X and Y are two sets, I implicitly assume that $Y\subset X$.

An *alphabet* is a finite set whose elements are called *letters*. The alphabet is usually denoted by A. A *finite word* (or for short, a *word*) on A is a finite sequence of elements of the alphabet. A word will be denoted by writing its letters one after the other. Unless otherwise stated, every word, and every set of words we will talk about, is implicitly on an alphabet denoted by A. The *length* of a word u, denoted by $|u|$, is the number of its letters. There is a word of length 0 which is denoted by ε. The number of occurrences of a letter α in a word u is denoted by $|u|_\alpha$. It is clear that $\sum_{\alpha\in A}|u|_\alpha=|u|$.

The *concatenation* of two words u and v, denoted by uv, is the word obtained by writing the letters of u and then those of v. A *factor* of a word u is a word v such that there exist words w and z such that $u=wvz$. A factor v of u is *proper* if it is different from u.

A word v is a *prefix* of a word u iff there exists a word w such that $u=vw$. It is a *suffix* of u iff there exists a word w such that $u=wv$. (Note: prefixes and suffixes are, respectively, called 'left factor' and 'right factor' by some authors.)

A *language* is a set of words. If L is a language, then L^* (respectively L^+) denotes the set of the concatenation of at least 0 (resp. 1) words in L. The set of all the words on an alphabet A, that is A^*, with the concatenation product is the free monoid on A. Note that A^+ is the set of the words different from ε. The set of the words on A of length l is denoted by A^l and the set of the words of length less than or equal to l is denoted by $A^{\leqslant l}$. Sometimes, the language $\{u\}$ will be denoted simply by u.

A *bi-infinite word* is a $\mathbb{Z}$-sequence of elements in A (An *infinite word* is an $\mathbb{N}$-sequence). The set of all bi-infinite words on A is denoted by $A^{\mathbb{Z}}$. A bi-infinite word is denoted by $(a_i)_{i\in\mathbb{Z}}$ or by $\ldots a_{-p}\, a_{-p+1}\ldots a_{-2}a_{-1}a_0\,a_1\,a_2\ldots a_q\ldots$

A *finite factor of a bi-infinite word* $\aleph=(a_i)_{i\in\mathbb{Z}}$ is a finite word u such that there are integers N and N' such that $u=a_N\ldots a_{N'-1}$.

If the sequence is periodic of period T, then the bi-infinite word will be said to be *periodic of period T*. Such a word $\aleph=(a_i)_{i\in\mathbb{Z}}$ is given by $(a_i)_{i\in\{0,\ldots,T-1\}}$. That word will be denoted by $u^{\mathbb{Z}}$ where $u=a_0\ldots a_{T-1}$.

Let $\aleph_1=(a_{1,n})_{n\in\mathbb{Z}}$ and $\aleph_2=(a_{2,n})_{n\in\mathbb{Z}}$ be two bi-infinite words on the alphabet A. They will be said to be *translates of each other* iff there is a p such that $\forall n\in\mathbb{Z}$, $a_{1,n}=a_{2,n+p}$. The notation $\aleph_1\equiv\aleph_2$ will be used to say $\aleph_1$ and $\aleph_2$ are translates of each other, and $\aleph_1\not\equiv\aleph_2$ to say they are not. The relation $\equiv$ is clearly an equivalence relation. From now on, bi-infinite words will always be considered up to translation. Note that if p,s are words (not both equal to ε), then $(ps)^{\mathbb{Z}}\equiv(sp)^{\mathbb{Z}}$.

An *infinite word* is a $\mathbb{N}$-sequence of elements in A. An infinite word can be denoted by $(a_i)_{i\in\mathbb{N}}$. If the sequence is periodic of period T, then the infinite word will be denoted by $u^{\mathbb{N}}$, where $u=a_0\ldots a_{T-1}$.

Let A be an alphabet, an *automaton* on A is a pair $(\mathcal{Q},F)$ where $\mathcal{Q}$ is a finite set called the set of *states*, and where F is a subset of $\mathcal{Q}\times A\times\mathcal{Q}$. The elements of F are called the *arrows* of the automaton. The notation $p\xrightarrow{a}q$ will express that the arrow (p,a,q) exists (In a nutshell, an automaton is a graph with letters on the arrows.). Note that I do not need to introduce initial and final states as is usually done in automata theory.

A finite word $u=a_1\ldots a_n$ is *recognized* by the automaton $(\mathcal{Q},F)$ iff there are states $q_0,\ldots,q_n$ in $\mathcal{Q}$ such that $\forall i$, $q_{i-1}\xrightarrow{a_i}q_i$. Such a sequence of arrows is called a *path* recognizing the word u and the notation $q_0\xrightarrow{u}{}^{*}\,q_n$, will express that such a path exists from q_0 to q_n. As a convention, $\forall p\in\mathcal{Q}$, $p\xrightarrow{\varepsilon}{}^{*}\,p$ always holds. A loop is a path $q\xrightarrow{u}{}^{*}\,q$ (which starts and ends at the same state). A bi-infinite word $\aleph=(a_i)_{i\in\mathbb{Z}}$ is *recognized* by the automaton $(\mathcal{Q},F)$ iff $\exists(q_i)_{i\in\mathbb{Z}}\in\mathcal{Q}$ such that $\forall i$, $q_{i-1}\xrightarrow{a_i}q_i$. That sequence of arrows is called a *bi-infinite path* recognizing the word $\aleph$. Note that if ω is a finite or bi-infinite word recognized by the automaton $\mathcal{A}$, then every finite factor of ω is also recognized by $\mathcal{A}$.

3. Unavoidable languages

3.1. Unavoidable languages: definition

Proposition 3.1. *Let X be a language, then the following two properties are equivalent*:

(i) *There is an integer N such that, for every word u in A^* of length at least N, there is a word v in X which is a factor of u.*

(ii) *For every word $\aleph$ in $A^{\mathbb{Z}}$, there is a word v in X which is a factor of $\aleph$. Moreover, if X is finite, then the above two properties are equivalent to the following one*:

(iii) *For every* periodic *word $\aleph$ in $A^{\mathbb{Z}}$, there is a word v in X which is a factor of $\aleph$.*

Proof. (i) $\Rightarrow$ (ii): Assume $\neg$(ii): There is a bi-infinite word $\aleph$ such that no element in X is a factor of $\aleph$. Let E be the set of the finite factors of $\aleph$. Elements in X are factors of *no* word in E. This language E contains words of every length and therefore (i) is not satisfied. One has $\neg$(i).

(ii) $\Rightarrow$ (i): Assume $\neg$(i), then there is, for every $n \in \mathbb{N}$, a word u_n with no factor in X and which is of length $2n+1$. Let $(a_{n,i})_{n\in\mathbb{N}, i\in\mathbb{Z}, -n\leqslant i\leqslant n}$ be the letters such that $u_n = a_{n,-n}a_{n,-n+1}\ldots a_{n,0}\ldots a_{n,n}$ for every $n \in \mathbb{N}$. Define $(a_{n,i})_{n\in\mathbb{N}, i\in\mathbb{Z}, |i|>n}$ in an arbitrary way. Then let $\aleph_n = (a_{n,i})_{i\in\mathbb{Z}}$ for every $n \in \mathbb{N}$. Put the discrete topology on the alphabet A, which becomes a compact metric space, and the infinite-product topology on $A^{\mathbb{Z}}$ which becomes also a compact metric space. Thus, one can extract from $(\aleph_n)_{n\in\mathbb{N}}$ a subsequence $(\aleph_{\phi(n)})_{n\in\mathbb{N}}$ which converges to a bi-infinite word $\aleph = (a_i)_{i\in\mathbb{Z}}$. The convergence of $(\aleph_{\phi(n)})_{n\in\mathbb{N}}$ to $\aleph$ implies that each finite factor of $\aleph$ is a factor at the same position of all but finitely many $\aleph_{\phi(n)}$'s, therefore is a factor of a word u_n for some $n \in \mathbb{N}$, and therefore is not in X. So no element in X is a factor of $\aleph$. Consequently, (ii) is not satisfied: One has $\neg$(ii).

(ii) $\Rightarrow$ (iii) is obvious.

If X is finite, then (iii) $\Rightarrow$ (i):

Assume (iii). Let $l = \max_{v\in X}|v|$, $K = (\operatorname{card} A)^l$ (K is the number of words on A of length l) and $N = (K+1)l$. Let u be a word of length at least N. The word u can be written $u = u_0u_1\ldots u_K\, z$ where for every i in $[0,k]$, u_i is a word of length l, and where z is a word. Because there are $K+1$ u_k's and only K different words of length l, two u_k's must be equal, i.e. $\exists i < j$ such that $u_i = u_j$. Let $w = u_iu_{i+1}\ldots u_{j-1}$. Because (iii) is assumed to be true, there is an $x \in X$ which is a factor of $w^{\mathbb{Z}}$. We have now two cases:

- If x is a factor of w, then x is also a factor of u since w is a factor of u.
- If x is not a factor of w, then there are an $n \in \mathbb{N}$, a suffix s and a prefix p of w such that $x = sw^np$. But $|w| = (j-i)l \geqslant l = \max_{v\in X}|v| \geqslant |x|$, therefore n must be 0 and $x = sp$ (or $n = 1$ and $s = p = \varepsilon$, but then $x = w$ which cannot happen here since

we have assumed that x is not a factor of w). But p is a prefix of $w=u_i\dots u_{j-1}$ and (since $x=sp$), $|p|\leqslant|x|\leqslant l=|u_i|$, therefore p is a prefix of u_i, which is the same as u_j. Since s is a suffix of $w=u_i\dots u_{j-1}$, since p is a prefix of u_j and since $x=sp$, one gets that x is a factor of $wu_j=u_i\dots u_{j-1}u_j$ which is a factor of u. Therefore x is a factor of u.

In both cases, x is found to be a factor of u, and (i) is proved.

Proposition 3.1 is proved. □

Notes 3.2.

- The implication (iii)⇒(i) is false for infinite languages, see for example $X=\{uu\mid u\in A^+\}$, the set of non-ε squares on the alphabet $A=\{a,b,c\}$ with the help of [2].
- When X is finite, another way to prove Proposition 3.1 is to build an automaton recognizing finite and infinite words with no factor in X and then to see that the above three conditions are equivalent to "there are no loops in the automaton".

Definitions 3.3. A language X is *unavoidable* iff it satisfies the first two conditions in Proposition 3.1, it is *avoidable* iff it does not.

Equivalent definitions are:

Let X be a language, then X is unavoidable iff:

- $A^*-A^*XA^*$ is finite: all but finitely many finite words have a factor in X.
- $A^{\mathbb{Z}}-A^{-\mathbb{N}}XA^{\mathbb{N}}$ is empty: all bi-infinite words have a factor in X.

Definition 3.4. Let X be a language and ω be a finite or a bi-infinite word, then ω *avoids* X if no element in X is a factor of ω.

Finite unavoidable languages are quite representative of unavoidable languages thanks to the following proposition:

Proposition 3.5. *Let X be an (infinite) unavoidable language. There is a finite sublanguage X' of X which is unavoidable.*

Proof. This proposition is proved by W. Bucher, A. Ehrenfeucht and D. Haussler in [3] and by C. Choffrut and K. Culik in [4]. A short proof of this fact is: Let S_w be the set of bi-infinite words containing w as a factor, then $\bigcup_{w\in X}S_w=A^{\mathbb{Z}}$. But with the infinite-product topology, the S_w's are open and $A^{\mathbb{Z}}$ is compact, thus there is a finite sublanguage X' of X such that $\bigcup_{w\in X'}S_w=A^{\mathbb{Z}}$, that is, which is unavoidable. Proposition 3.5 is proved. □

3.2. Examples of unavoidable languages

- $X=A$ is unavoidable.
- $\forall n\in\mathbb{N}$, $X=A^n$ (The set of the words of length n) is unavoidable.
- If $A=\{a,b\}$, then $X=\{aa,bab,bbbbbbbbbb\}$ is unavoidable.

Indeed, try to construct a bi-infinite word $\aleph$ which avoids X: all a's must be preceded and followed by a b because $aa \in X$, thus must be included in a factor bab. But $bab \in X$, so there must be no a's in $\aleph$, so $\aleph$ has to be $b^{\mathbb{Z}}$, but then it contains $bbbbbbbbbb$ which is in X. Thus no bi-infinite word can avoid X, which is therefore unavoidable.

- If $A = \{a, b\}$, then $X = \{bb, bab, baab, baaab, \ldots, ba^i b, \ldots, ba^n b, ba^{n+1}, a^{n+2}\}$ is unavoidable.

Indeed, assume $\aleph$ is a bi-infinite word which avoids X and contains a b. This b must be followed by an a, because bb is in X. Moreover, there cannot be more than $n+1$ consecutive a's after that b because ba^{n+1} is in X. Thus, after that b, there are k a's where $1 \leqslant k \leqslant n$, and those a's are followed by a b, so that $\aleph$ contains $ba^k b$, which is in X, there is a contradiction. Thus there cannot be any b in a word avoiding X, so there are only a's, but this is also forbidden, since a^{n+2} is in X. So, no bi-infinite word avoids X, which is therefore unavoidable.

- If $X = \{l_1, \ldots, l_n\}$, if $X' = \{l'_1, \ldots, l'_n\}$, if l_i is a factor of l'_i for all i and if X' is unavoidable, then X is also unavoidable.
- If $X \subset X'$ and if X is unavoidable, then X' is also unavoidable.

By looking at the last example, one can see that unavoidable languages can be uselessly big. The minimal unavoidable languages will be defined in Section 5.

3.3. *The unavoidable languages of n words on $A = \{a, b\}$, n being small*

- $n = 0$: nothing
- $n = 1$: $X = \{\varepsilon\}$ is the only one-element unavoidable language.
- $n = 2$: The unavoidable languages of two words are:
  - The languages of two words containing ε
  - $\{\{a, b^n\} \mid n \in \mathbb{N}\}$
  - $\{\{b, a^n\} \mid n \in \mathbb{N}\}$

Indeed, all these languages are unavoidable, and a language of two words which is unavoidable and which does not contain ε, must contain at least one factor of each of the periodic bi-infinite words $a^{\mathbb{Z}}$, $b^{\mathbb{Z}}$ and $(ab)^{\mathbb{Z}}$. This is possible only if the language is in the above list.

- $n = 3$: The unavoidable languages of three words are:
  - The languages of three words containing an unavoidable language of 1 or 2 words, which were previously described
  - $\{\{aa, b^n, bab\} \mid n \in \mathbb{N}\}$
  - $\{\{bb, a^n, aba\} \mid n \in \mathbb{N}\}$
  - $\{\{aa, bb, u\} \mid u \in (\varepsilon + b)(ab)^*(\varepsilon + a)\}$
  - $\{\{a^m, b^n, u\} \mid m,\ n \in \mathbb{N},\ u \in \{ab, ba\}\}$

(Note: In fact, in the last case, one should specify $(m, n) \neq (0, 0)$, otherwise $\{a^0, b^0, u\} = \{\varepsilon, u\}$ is not a language of three elements.)

Indeed, all these languages are unavoidable (for the latter one, a word avoids ab iff it is in b^*a^* and every word in b^*a^* of length greater than $m+n$ contains a^m or b^n),

and every unavoidable language of three words is in the above list because:

If X is such a language, then X contains at least one factor of each of the words $\aleph_1 = a^{\mathbb{Z}}$, $\aleph_2 = b^{\mathbb{Z}}$, $\aleph_3 = (ab)^{\mathbb{Z}}$ and $\aleph_4 = (aab)^{\mathbb{Z}}$. Since X contains three words and there are four $\aleph_i$'s, there must be a word in X which is a factor of at least two $\aleph_i$'s, therefore X must contain one of the words $\varepsilon, a, b, aa, ab, ba, aba$.

- If X contains ε, then it is in the above list.
- If X contains a, then: The language X contains also a factor of $b^{\mathbb{Z}}$, that is it contains b^m for some integer m, and therefore it contains $\{a, b^m\}$ and it is in the above list.
- If X contains b, then: The language X contains also a factor of $a^{\mathbb{Z}}$, that is it contains a^m for some integer m, and therefore it contains $\{b, a^m\}$ and it is in the above list.
- If X contains ab or ba (and not ε), then: It contains also a factor of $a^{\mathbb{Z}}$ and one of $b^{\mathbb{Z}}$, so is element of $\{\{a^m, b^n, u\} \mid m, n \in \mathbb{N},\ u \in \{ab, ba\}\}$ and it is therefore in the above list.
- If X contains aa, then: It contains also a factor of each of the words $\aleph_1 = b^{\mathbb{Z}}$, $\aleph_2 = (ab)^{\mathbb{Z}}$ and $\aleph_3 = (abb)^{\mathbb{Z}}$. So X contains a word which is a factor of at least two $\aleph_i$'s (recall that there are two words in $X - \{aa\}$), so X contains one of the words $\varepsilon, b, bb, ab, ba, bab$.
  (1) If X contains ε, b, ab or ba, then it has already been considered in the previous cases.
  (2) If X contains bb, then: It contains also a factor of $(ab)^{\mathbb{Z}}$, it is therefore element of $\{\{aa, bb, u\} \mid u \in (\varepsilon + b)(ab)^*(\varepsilon + a)\}$ and it is described in the above list.
  (3) If X contains bab, then: It contains also a factor of $b^{\mathbb{Z}}$, it is therefore one of the $\{\{aa, b^n, bab\} \mid n \in \mathbb{N}\}$ and it is described in the above list.
- If X contains aba, then: It contains also a factor of each of the words $\aleph_1 = a^{\mathbb{Z}}$, $\aleph_2 = b^{\mathbb{Z}}$ and $\aleph_3 = (abb)^{\mathbb{Z}}$, thus it contains one of the words ε, a, b, bb and it has therefore already been considered in the previous cases.

4. Inventories

In Section 3.3, I made the inventories of unavoidable languages of n words where $n = 0, 1, 2, 3$. The aim of this section is to prove that one can make such an inventory for every $n \in \mathbb{N}$.

First, I will define descriptions which will be used to make the inventories, then I will prove that the set of unavoidable languages of a given cardinality is describable, and to finish with, I will give some extras explanations on how to compute explicitly the descriptions.

First, let us introduce a notation:

Notation 4.1. Let u be a word, $u^\bullet$ denotes the set of the factors of $u^{\mathbb{Z}}$.

Note that if $u = a_1 \ldots a_n$, then $u^\bullet = (\varepsilon + a_n + \cdots + a_2 \ldots a_n)u^*(\varepsilon + a_1 + \cdots + a_1 \ldots a_{n-1}) + f(u)$, where $f(u)$ is the set of the factors of u (to catch words such as

$a_2 \dots a_{n-1}$). For example, one has $a^\bullet = a^*$, $(ab)^\bullet = (\varepsilon + b)(ab)^*(\varepsilon + a)$ and $(abcd)^\bullet = (\varepsilon + d + cd + bcd)(abcd)^*(\varepsilon + a + ab + abc) + b + c + bc$.

Now, we can introduce the descriptions:

Definitions 4.2. Let N be an integer. An *elementary N-description* is a p-uple, where p is an integer less than or equal to N, of languages $R_1 * \cdots * R_p$, where for every i in $[1, p]$, either R_i is a singleton or there is $u \in A^+$ such that $R_i = u^\bullet$.

This elementary N-description *describes* the set of finite languages of at most N words: $\{\{w_1, \dots, w_q\} \mid p \leqslant q,\ w_1 \in R_1, \dots, w_p \in R_p\}$.

A set of languages is *elementarily describable* if there is an elementary description which describes it.

Definitions 4.3. An *N-description* is a finite set of elementary N-descriptions. It *describes* the union of the sets of the languages described by the elementary N-descriptions. A set of languages is *describable* if there is a description which describes it.

Note 4.4. A language is described by an elementary description iff it contains at most N elements and contains a sublanguage $\{\{w_1, \dots, w_p\} \mid w_1 \in R_1, \dots, w_p \in R_p\}$; Note that such a language may have less than p words and that q might be more than N, for example, $R_1 = a^\bullet$, $R_2 = (ab)^\bullet$ describes $\{a\}$ with $\{N = p = q = 2,\ w_1 = w_2 = a\}$ and $\{a, bb\}$ with $\{N = p = 2,\ q = 3,\ w_1 = w_2 = a,\ w_3 = bb\}$. I consider these situations to be weird, but it turns out to be more convenient to allow such side-effects than to forbid them. The elements $\{w_{p+1}, \dots, w_q\}$ are not relevant, they are useless extra words. Here again, I added them so that the statements of this section are kept correct without adding useless fussing details.

As an example, the lists made in previous section show that the unavoidable languages of at most N words on the alphabet $\{a, b\}$ are describable if $N \leqslant 3$. This is the statement I intend to prove for every N.

Theorem 4.5. *Let A be an alphabet and N an integer, the set of unavoidable languages of at most N words on A is N-describable.*

Proof. The idea is the following:

To be unavoidable, a language X of N words cannot contain only "long words", that is, it must contain a word in a finite language $L_\emptyset$ that one can calculate.

For every choice of w in $L_\emptyset$, one looks for unavoidable languages of N words that contain w and one (usually) sees that they must contain a word in another finite language L_w.

For every choice of w' in L_w, one looks for unavoidable languages of N words that contain w and w' ...

It will be convenient to introduce the following definitions:

Definition 4.6. Let Y be a language. A *completion of Y into an unavoidable language* (or for short, a *completion of Y*) is a finite language Z such that $Y \cap Z = \emptyset$ and such that $Y \cup Z$ is unavoidable. An *M-completion of Y (into an unavoidable language)* is a completion Z of Y, whose cardinality is at most M.

Let Y be a language and M be an integer, the set of all the unavoidable M-completions of Y (or for short the *M-completion set of Y*) will be denoted by $\mathscr{C}_{Y,M}$.

It is for the sake of convenience that it is required that a completion satisfies $Y \cap Z = \emptyset$, and that completions of cardinality less than M are included among M-completions.

Theorem 4.5 is an obvious corollary of the following key lemma (by considering $Y = \emptyset$ and $M = N$):

Lemma 4.7. *Let A be an alphabet, M an integer and Y a finite language on A, then $\mathscr{C}_{Y,M}$ is M-describable.*

(The idea is that M is the number of words that remain to be found when one looks for unavoidable languages of N words containing Y.)

Proof. This lemma is proved by induction on M.

- $M = 0$: according to whether Y is unavoidable or not, $H_{Y,0}$ will be the singleton $\{\emptyset\}$ (described by one elementary 0-description which is empty) or empty (described by zero elementary 0-description).
- Assume the result is true for $M-1$ and for every finite language Y.

  Let $Y = \{w_1, \ldots, w_n\}$ be a finite language. One proves the result for Y and M.

  There are two cases:

  Case 1: There are at least $M+1$ periodic bi-infinite words $(\aleph_i)_{0 \leqslant i \leqslant M}$, which are different up to translation, and such that for every i in $[0, M]$, the word $\aleph_i$ has no factor in Y.

  Let $F_{(\aleph_i)_{i \in [0,M]}}$ be the set of the words which are factor of at least two different $\aleph_i$'s (Note that the intersection with Y is empty because all the elements in $F_{(\aleph_i)_{i \in [0,M]}}$ are factor of some word $\aleph_i$, and none in Y is).

  Every M-completion Z of Y has an element in $F_{(\aleph_i)_{i \in [0,M]}}$. Indeed, assume Z is an M-completion of Y. Since $Y \cup Z$ is unavoidable, each word $\aleph_i$ has a factor in $Y \cup Z$, but no $\aleph_i$ has a factor in Y, thus each $\aleph_i$ has a factor in Z. There are $M+1$ words $\aleph_i$ while the cardinal of Z is M, so there are some integers $i, j \in [\![0, n]\!]$ with $i \neq j$ such that $\aleph_i$ and $\aleph_j$ have a common factor w in Z.

  Note that if two different periodic words (of periods T_1 and T_2) have a common factor, then the length of this common factor is bounded (it is less than $T_1 + T_2$). (Proof left to the reader or see [8, 15]), so that $F_{(\aleph_i)_{i \in [0,M]}}$ is finite.

  Let w be in $F_{(\aleph_i)_{i \in [0,M]}}$. The M-completions of Y containing w are exactly the $Z' \cup \{w\}$, Z' being an $(M-1)$-completion of $Y \cup \{w\}$. Those $(M-1)$-completions are

describable (induction hypothesis) and therefore, the M-completions of Y containing w are describable (Add $R=\{w\}$ to each elementary $(M-1)$-description of $(M-1)$-completion of $Y\cup\{w\}$).

Since $F_{(\aleph_i)_{i\in[0,M]}}$ is finite, M-completions of Y ($=M$-completions of Y containing a word in $F_{(\aleph_i)_{i\in[0,M]}}$) are describable. The description is just obtained by taking the union on $w\in F_{(\aleph_i)_{i\in[0,M]}}$ of the descriptions of the completions containing $\{w\}$.

Case 2: There is an integer $m\leqslant M$ such that there are only m periodic bi-infinite words $(\aleph_i)_{1\leqslant i\leqslant m}$ which are different up to translation and such that no word $\aleph_i$ has a factor in Y.

Let $(u_i)_{1\leqslant i\leqslant m}$ be words such that $\aleph=u_i^{\mathbb{Z}}$. If Z is a language, then $Y\cup Z$ is unavoidable iff for all i with $1\leqslant i\leqslant m$, there is a word z_i in Z such that $z_i\in u_i^{\bullet}$. Therefore M-completions of Y are described by $(R_i)_{i\in[1,m]}$, where $R_i=u_i^{\bullet}$ for every $i\leqslant m$. Note that this includes the M-completions Z with a word $z\in Z$ which is factor of at least two words $\aleph_i$.

Therefore, the M-completion set of Y is describable.

The result is true for M, and thus, by induction, for every $M\in\mathbb{N}$.

Lemma 4.7 and therefore Theorem 4.5 are proved. □

The proof in last section is a recursive algorithm provided one gets a way to figure out whether there are more than $M+1$ periodic bi-infinite words that avoid a given language Y, and if yes, to get explicitly $M+1$ of these words, if not, to get all the periodic bi-infinite words avoiding Y.

The first step to do that is to build an automaton which recognizes the set A_Y of the words with no factor in Y. Since Y is finite, it is obvious that A_Y is regular. Moreover, there is an automaton which is linear in size (the number of state is less than or equal to $1+\sum_{y\in Y}|y|$) and which can be built in linear time. The base of this automaton is the one that A.V. Aho and M.J. Corasick introduced in [1], and which is already used by Choffrut and Culik in [4].

Without more explanations, I here give the construction of the automaton I need:

Assume Y is a finite language such that no element in Y is a factor of another one (that is, $\forall y,y'\in Y$, $[y\neq y'\Rightarrow y$ is not a factor of $y']$).

(Note: the languages Y which are considered during the calculation of the set of unavoidable languages of given cardinality will satisfy this property.)

Let $\mathscr{A}$ be the following automaton:

- The set of states $\mathscr{Q}$ is the set of the prefixes of words in Y.
- For each $u\in\mathscr{Q}$ and $a\in A$, there is an arrow $u\xrightarrow{a}v$ where v is the longest suffix of ua in $\mathscr{Q}$.

It is left to the reader to prove that every bi-infinite word is recognized by a unique bi-infinite path, and that a word $\aleph$ avoids a word $y\in Y$ iff the corresponding path does not go through the state y. Therefore, bi-infinite words avoiding Y are recognized by the automaton $\mathscr{B}$ defined by

- The states of $\mathscr{B}$ are the states of $\mathscr{A}$ which are not elements of Y.
- The arrows of $\mathscr{B}$ are the arrows of $\mathscr{A}$ whose both ends are states of $\mathscr{B}$.

Since we are considering only bi-infinite periodic words, we can clean up the automaton and keep only the arrows of strong connected components, that is consider the automaton $\mathscr{C}$ whose states are the states of $\mathscr{B}$ and whose arrows are the arrows $p \xrightarrow{a} q$ of $\mathscr{B}$ such that $q \xrightarrow{*}_{\mathscr{B}} p$.

The bi-infinite periodic words which are recognized by C are those which avoid Y, and it is not very difficult to know from the automaton if there are infinitely many such words and to explicit (finitely many of) them

5. Minimal and reduced unavoidable languages

This section answers a question by C. Choffrut.

5.1. Dickson's lemma

Definitions 5.1. (1) The partial order $\leqslant$ is defined on $\mathbb{N}^n$ by

$$[(x_1,\ldots,x_n)\leqslant(y_1,\ldots,y_n)] \text{ iff } [x_1\leqslant y_1, x_2\leqslant y_2,\ldots,x_n\leqslant y_n].$$

(2) A subset S of $\mathbb{N}^n$ is an *ideal* iff

$$\forall \bar{x},\bar{y}\in\mathbb{N}^n,\quad [(\bar{x}\in S \text{ and } \bar{x}\leqslant\bar{y})\Rightarrow \bar{y}\in S].$$

(3) An element $\bar{x}$ of a subset S of $\mathbb{N}^n$ is *minimal* iff

$$\forall \bar{y}\in\mathbb{N}^n,\quad (\bar{y}\leqslant\bar{x} \text{ and } \bar{y}\in S)\Rightarrow \bar{y}=\bar{x}.$$

In dimension 2, S is an ideal iff for every $(x,x')\in S$, the quarter of the plane placed northeast of (x,x') is in S. When drawing an ideal, the minimal elements are the ones which are at a corner.

Examples 5.2. Some ideals in dimension 2 and 3 are to be found in Fig. 1. On the one in dimension 2, (1,9), (2,6), (3,5), (6,3) and (9,2) are the minimal elements in S.

The following was proved by Dickson in [6].

Proposition 5.3. *Let S be a ideal of $\mathbb{N}^n$, then the number of minimal elements in S is finite.*

Proof. It is proved by induction on n.

- $n=0$: obvious.
- Let us assume the result is true for ideals of dimension n.

Let S be an ideal of dimension $n+1$. Let $p=(x_1,\ldots,x_n,x_{n+1})$ be a minimal element in S (If there is no minimal element, then the result is true. This happens only with $S=\emptyset$.)

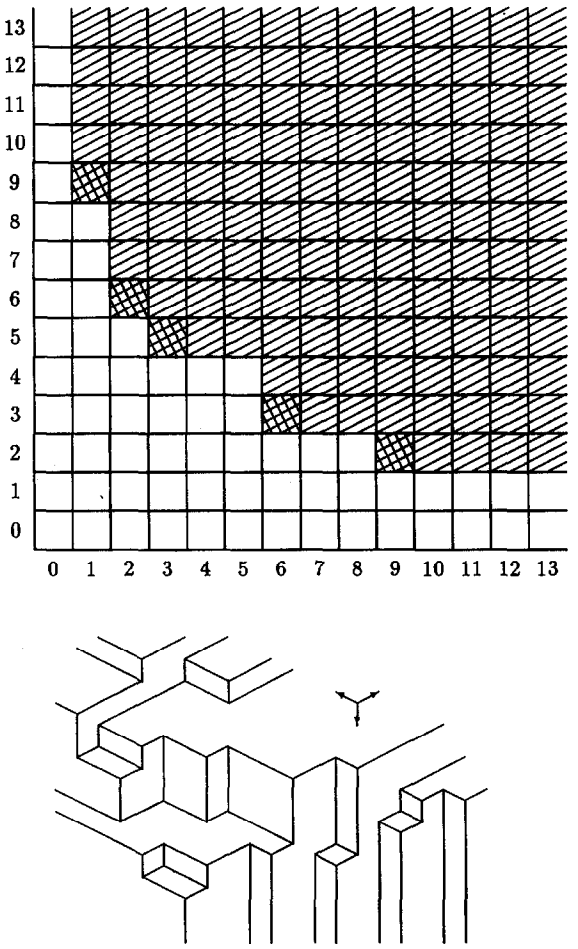

Fig. 1. Some ideals in dimensions 2 and 3.

An element in S greater than p (that is an element $q=(y_1,\ldots,y_{n+1})$ with $y_i \geqslant x_i$ for every i) is not a minimal element unless it is p itself. So if $q=(y_1,\ldots,y_{n+1})$ is a minimal element different from p, then there is an i such that $y_i < x_i$.

For every $i \in [1,n+1]$ and $y < x_i$, we define the *the projection of S according to* (i,y) as:

$$\{(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_{n+1}) \mid (x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_{n+1}) \in S\}.$$

The reader can check that the projection of an ideal of dimension $n+1$ is a ideal of dimension n, and that images of minimal elements are minimal elements (but not inverse images).

So, for each choice of i and $y < x_i$ (And there are only finitely many such choices), the minimal elements in S with $y_i = y$ are minimal elements of an ideal of dimension

n and are, thanks to the induction hypothesis, finite in number. The set of minimal elements in S is the finite union of the set of minimal elements with $y_i = y < x_i$ and of $\{p\}$, so is finite. So the result is true for $n+1$.

• By induction, the result is true for all $n \in \mathbb{N}$, so Proposition 5.3 is proved. □

Note 5.4 (*due to D. Perrin*). There is a funny way to prove this lemma by using unavoidable languages:

Ehrenfeucht and Rozenberg proved in [7] the following:

Let X be a language, and let u and v be two words. Let $\leqslant_X$ be the reflexive and transitive closure of the relation $\leqslant_1$ defined by $u \leqslant_1 v$ iff $\exists w, y, z$ with $w, z \in A^*$ and $y \in X$, such that $u = wz$ and $v = wyz$. (In a nutshell, $u \leqslant_X v$ iff u can be obtained from v by successive deletions of factors which are elements of X.) It is easy to see that $\leqslant_X$ is a partial order relation.

The theorem is:

There is an infinite language Y such that $\forall u, v \in Y$, $u \neq v \Rightarrow u \not\leqslant_X v$ and $v \not\leqslant_X u$ iff X is NOT unavoidable.

Let $X = A$, then $\leqslant$ is the order "is a sub-word of". But A is unavoidable, so according to the theorem, there is no infinite language with no $u, v, u \neq v$, u sub-word of v. This is known as Higman's theorem [9].

Now, let C be the set of the minimal elements of an ideal S of $\mathbb{N}^n$. Let f be defined from $\mathbb{N}^n$ to A^n (where $A = (a_1, \ldots, a_n)$) by $f(x_1, \ldots, x_n) = a_1^{x_1} \ldots a_n^{x_n}$. One has $f(u) \leqslant f(v) \in A^*$ iff $u \leqslant v \in \mathbb{N}^n$. So $f(C)$ is a set of incomparable words, so is finite, so, (since f is injective) C is finite. □

Note 5.5. There is a notion of well partial ordering relations, which has several equivalent definitions. Dickson's lemma says that the order defined on $\mathbb{N}^n$ is a well partial ordering relation and the proof of the note 1 can be rewritten: $\leqslant_X$ is a well partial ordering relation iff X is unavoidable, therefore $\leqslant_A$ is a well partial ordering relation, which implies that the relation $\leqslant$ on $\mathbb{N}^n$ is a well partial ordering relation. See [16] for definitions of well partial ordering relations and for details on the well partial ordering relation $\leqslant_A$.

5.2. Minimal and reduced unavoidable languages: definitions

If Y is a subset of X and if Y is unavoidable, then X is also unavoidable, but $X - Y$ is unnecessary to make X unavoidable. An unavoidable language is minimal iff there is no unnecessary word, i.e.

Definition 5.6. Let X be an unavoidable language, X is *minimal* iff every proper subset Y of X (every language Y such that $Y \subset X$ and $Y \neq X$) is not unavoidable.

Definition 5.7. Let X and Z be two languages, Z is *a reduction of* X iff there is an integer r, distinct words $(z_i)_{1 \leqslant i \leqslant r}$ and distinct words $(x_i)_{1 \leqslant i \leqslant r}$ such that $Z = \{z_1, \ldots, z_r\}$, $X = \{x_1, \ldots, x_r\}$ and $\forall i$, z_i is a factor of x_i.

Remark 5.8. If Z is a reduction of X, if Z is minimal unavoidable and if X is unavoidable, then X is minimal.

Definition 5.9. Let X be a minimal unavoidable language, X is *reduced* iff X is minimal unavoidable and no proper reduction of X is minimal, which is in fact equivalent to:
$\forall wa \in X,\ X - \{wa\} + \{w\}$ is not minimal, and
$\forall aw \in X,\ X - \{aw\} + \{w\}$ is not minimal.

One can easily show that if X is unavoidable, then there is a subset Y of X which is minimal unavoidable (even if X is infinite, thanks to Proposition 3.5), and that if Y is minimal unavoidable, then there is a reduction Z of Y which is reduced.

5.3. The number of reduced unavoidable languages

Theorem 5.10. *Let A be an alphabet and N be an integer. There is a finite number of reduced unavoidable languages of N words on A.*

Proof. The proof will use the following definition:

Definition 5.11. An unavoidable language X of N words is *reduced in the elementary description $\mathscr{E}$* iff X is described by $\mathscr{E}$, X is minimal and for every language Y:

$$\left.\begin{array}{l} Y \text{ is described in } \mathscr{E} \\ Y \text{ is a reduction of } X \end{array}\right\} Y \Rightarrow \text{ is not minimal.}$$

If X is described in $\mathscr{E}$ and is reduced, then it is reduced in $\mathscr{E}$ (the converse is false). Since unavoidable languages of N words are describable by a finite number of elementary descriptions, it is sufficient to prove that for every elementary description of unavoidable languages $\mathscr{E}$, reduced elements in $\mathscr{E}$ are finite in number.

Let $\mathscr{E} = R_1 \times \cdots \times R_p$ be an elementary description of unavoidable languages of at most N words.

If $p < N$, then every language of N words described by $\mathscr{E}$ contains a proper sublanguage which is unavoidable. Therefore, no language of N words described by $\mathscr{E}$ is reduced in $\mathscr{E}$.

So it can be assumed that $p = N$ and it can also be assumed that there is an integer $n \leqslant N$, and some words $(w_i)_{1 \leqslant i \leqslant n}$ and $(u_i)_{n \leqslant i \leqslant N}$, such that:

$$R_1 = \{w_1\},\ \ldots,\ R_n = \{w_n\},\quad R_{n+1} = u_{n+1}^{\bullet},\ \ldots,\ R_N = u_N^{\bullet}.$$

Lemma 5.12. *Let $z \in u^{\bullet}$, then there are words s, t, p and an integer i such that p is a prefix of st, $u = ts$, and $z = (st)^i p$.*

Proof. The proof of this lemma is left to the reader. □

Let $\mathscr{F} = (s_l, t_l, p_l)_{n \leqslant l \leqslant N}$ be such that $u_l = t_l s_l$ and such that p_l is a prefix of $s_l t_l$.

$X=\{w_1,\ldots,w_n,(s_{n+1}t_{n+1})^{i_{n+1}}p_{n+1},\ldots,(s_Nt_N)^{i_N}p_N\}$ will be called reduced in $\mathscr{E}$ according to $\mathscr{F}$ iff it is minimal and $[j_{n+1}\leqslant i_{n+1},\ldots,j_N\leqslant i_N,(j_{n+1},\ldots,j_N)\neq(i_{n+1},\ldots,i_N)\Rightarrow X=\{w_1,\ldots,w_n,(s_{n+1}t_{n+1})^{j_{n+1}}p_{n+1},\ldots,(s_Nt_N)^{j_N}p_N\}$ is not minimal].

If X is reduced in $\mathscr{E}$, then it is reduced in $\mathscr{E}$ according to $\mathscr{F}$ (the converse is not true). As there are only finitely many choices for $\mathscr{F}$, it is sufficient to prove that for each choice of $\mathscr{F}$, reduced elements in $\mathscr{E}$ according to $\mathscr{F}$ are finite in number.

But $S=\{(i_{n+1},\ldots,i_N),\{w_1,\ldots,w_n,(s_{n+1}t_{n+1})^{i_{n+1}}p_{n+1},\ldots,(s_Nt_N)^{i_N}p_N\}$ is minimal$\}$ is an ideal (see Remark 5.8) and $X=\{w_1,\ldots,w_n,(s_{n+1}t_{n+1})^{i_{n+1}}p_{n+1},\ldots,(s_Nt_N)^{i_N}p_N\}$ is reduced in $\mathscr{E}$ according to $\mathscr{F}$ iff $(i_{n+1},\ldots,i_N)$ is a minimal element in S. Thus, according to Dickson's lemma, the number of reduced elements in $\mathscr{E}$ according to $\mathscr{F}$ is finite.

So the number of reduced elements in $\mathscr{E}$ is finite.

So the number of reduced unavoidable languages of N words is finite.

Theorem 5.10 is proved. □

6. The word-extension conjecture

6.1. The statement of the conjecture

The reader can check that if the language X is unavoidable, then $[wa\in X\Rightarrow(X-\{wa\})+\{w\}$ is also unavoidable]. (One can reduce words and stay unavoidable.). But if $w\in X$, then $X-\{w\}+\{wa\}$ is not unavoidable in general: one cannot extend X in any way and keep unavoidability. The question is: is there always a way to extend X and keep unavoidability, that is, is

$$P(X)\text{: }[X\text{ unavoidable}\Rightarrow\exists w\in X,\ \exists a\in A,\ X-\{w\}+\{wa\}\text{ is unavoidable}]$$

true?

It is not for $X=\{\varepsilon\}$, but no other counter-example was known.

It was conjectured that P is always true (except for $X=\{\varepsilon\}$).

Now, let us look at unavoidable sets of N words.

Let $\mathscr{E}$ be an elementary N-description of unavoidable languages. Then:

Either

$p<N$ (see the definition of descriptions for the meaning of p): Then every set Y of N words described by $\mathscr{E}$ contains at least one useless element w (i.e. a word w such that $Y-\{w\}$ is unavoidable) and one can obviously replace w by wa and keep unavoidability.

or

$\exists i$, $\exists u$ such that $R_i=u^\bullet$. Then $P(X)$ is true for all N-tuples described by $\mathscr{E}$.

or

$p=N$ and $\forall i,R_i$ is a singleton $\{w_i\}$: Then, one can test $P(X)$ for $\{w_1,\ldots,w_N\}$ which is the only language described by $\mathscr{E}$. (Note: this test can be made efficiently thanks to Aho and Corasick automaton [1].)

As unavoidable sets of N words are describable, then the word-extension conjecture is decidable for sets of at most N words, N being fixed.

It is especially interesting to look at unavoidable sets of seven words, for this leads to

6.2. A counterexample to the word-extension conjecture

Let $X = \{aaa, bbbb, abbbab, abbab, abab, bbaabb, baabaab\}$.

The language X is unavoidable.

One can give different proofs of this:

Proof 1, by hands:

Assume $\aleph$ is an infinite work that avoids X then:

If *bab* is a factor of $\aleph$, then *bab* cannot be preceded by an *a* (because *abab* is in X), so is preceded by a *b*, and *bbab* is a factor of $\aleph$, but then *bbab* cannot be preceded by an *a* (because *abbab* is in X), so is preceded by another *b* and *bbbab* is a factor of $\aleph$, but *bbbab* cannot be preceded by an *a* (because *abbbab* is in X), nor by a *b* (because then *bbbbab*, and therefore *bbbb* would be factors of $\aleph$ and *bbbb* is in X). Therefore *bab* is not a factor of $\aleph$.

So every *a* in $\aleph$ is preceded or followed by an *a*, so is included in a factor *aa*. This factor *aa* cannot be preceded nor followed by another *a* because then *aaa* (which is in X) would be a factor of $\aleph$, so the factor *aa* is preceded and followed by a *b*. Therefore, any factor *a* of $\aleph$ is included in a factor *baab*.

If *aba* is a factor of $\aleph$, then both the first and the last *a* of *aba* must be included in a factor *baab*, and therefore *aba* is itself included in a factor *baabaab*, but $baabaab \in X$, so *aba* cannot be a factor of $\aleph$.

Now any factor *a* of $\aleph$ is included in a factor *baab*, but *baab* cannot be preceded, nor followed by an *a*, because otherwise, $\aleph$ would contain the factor *aba*. So *baab* is followed and preceded by a *b* and $\aleph$ contains the factor *bbaabb*. But $bbaabb \in X$, so *a* cannot be a factor of $\aleph$. So $\aleph$ must be equal to $b^{\mathbb{Z}}$, but then it contains *bbbb* as a factor. This is impossible since *bbbb* is in X. So $\aleph$ cannot exist and X is unavoidable. □

Proof 2, with cuts: See [19] for the definition of cuts. Left to the reader or see [20].

Proof 3, with the automaton of X: Drawing the automaton of X and using it to check that X is unavoidable and not extendible is left to the reader.

The language X is counterexample to the conjecture, namely: $\forall x \in X$, $\forall u \in \{xa, xb, ax, bx\}$, $X - \{x\} + \{u\}$ *is not unavoidable.*

Indeed, let:

$$\aleph_{aaa} = a^{\mathbb{Z}} \qquad \aleph'_{aaa} = (baaa)^{\mathbb{Z}}$$

$$\aleph_{bbbb} = b^{\mathbb{Z}} \qquad \aleph'_{bbbb} = (bbbba)^{\mathbb{Z}}$$

$$\aleph_{abab} = (ab)^{\mathbb{Z}} \qquad \aleph'_{abab} = (aabab)^{\mathbb{Z}}$$

$$\aleph_{abbab} = (abb)^{\mathbb{Z}} \qquad \aleph'_{abbab} = (aabbab)^{\mathbb{Z}}$$

$$\aleph_{abbbab} = (abbb)^{\mathbb{Z}} \qquad \aleph'_{abbbab} = (aabbbab)^{\mathbb{Z}}$$
$$\aleph_{bbaabb} = (aabb)^{\mathbb{Z}} \qquad \aleph'_{bbaabb} = (aabbb)^{\mathbb{Z}}$$
$$\aleph_{baabaab} = (aab)^{\mathbb{Z}} \qquad \aleph'_{baabaab} = (baabaab)^{\mathbb{Z}}$$

The reader can check that for all $x \in X$, x is a factor of both $\aleph_x$ and $\aleph'_x$, but that for all $x, x' \in X$, $x \neq x'$, x' is not a factor of $\aleph_x$, nor of $\aleph'_x$. And if v is one the words xa, xb, ax or bx, then either $\aleph_x$ or $\aleph'_x$ avoids v, and since it avoids $X - \{x\}$, it avoids $X - \{x\} + \{v\}$ which therefore is NOT unavoidable. So $P(X)$ is not satisfied and the conjecture is not true. □

7. Open problems

Enlargement into a non-extendible unavoidable language

An extendible unavoidable language is an unavoidable language X such that there is $x \in X$ and some letters $a_1, a_2, \ldots, a_n, \ldots$ such that $X - \{x\} + \{x a_1 a_2 \ldots a_n\}$ is unavoidable for every integer n.

Open problem 7.1. Let Y be a finite language. What is a constructive necessary and sufficient condition on Y for there exists a completion of Y into a *non*-extendible unavoidable language?

A necessary condition is: For every $y \in Y$, there are at least two different periodic bi-infinite words $\aleph_y$ and $\aleph'_y$ such that y is a factor of both $\aleph_y$ and $\aleph'_y$, and such that y is the only element in Y to be a factor of $\aleph_y$ and the only one to be a factor of $\aleph'_y$.

This condition is not sufficient as one can see by considering the example $Y = \{ab\}$.

The word extension condition and reduced unavoidable languages

See Sections 5.2 and 6.

I gave in Section 6.2 an example of a language which does not satisfy the word-extension condition. This example is not reduced.

Open problem 7.2. Is it possible to find a reduced unavoidable language which does not satisfy the word extension condition or is this condition always satisfied by reduced unavoidable languages?

Bound for the longest word in a reduced unavoidable language

This problem was raised by Georges Hansel.

Let N be an integer (and A be an alphabet), there is a finite number of reduced unavoidable languages of N words, therefore one can speak about H_N the maximal length of the words which appear in reduced unavoidable languages.

Open problem 7.3. Give a bound for H_N. Is $H_N \leqslant N$?

Unavoidable sets in a monoid

I assume for this part that the reader knows about theory of semi-groups. See, e.g., [14] for undefined terminology.

One way to generalize unavoidable languages on a finite alphabet to unavoidable subsets of a monoid is the following:

Definition 7.4. Let S be a monoid and X a subset of S. The set X is unavoidable if there is a *finite* subset H of S such that for every element s in $S - H$, there are two elements u and v in S and an element x in X satisfying $s = uxv$. An unavoidable subset X of S is minimal if no proper subset of X is unavoidable.

It can easily be see that Definition 7.4 is equivalent to the usual definition if $S = A^*$ where A is a *finite* alphabet.

Open problems 7.5. Let R be a finite number of relations on A^*, where A is a finite alphabet, and let S be the monoid $A^*/\sim$, where $\sim$ is the congruence generated by R. Let X be a finite subset of S. Can one decide whether X is unavoidable? Does this have to do with the question whether equality is decidable in S? (If equality is decidable, is unavoidability decidable?, and conversely?) Does every unavoidable set contains a subset which is minimal unavoidable?, finite minimal unavoidable? Is it possible to describe unavoidable sets of a given cardinality? What if $\sim$ is a congruence generated by only one relation?...

Unavoidable sets among the square-free words

In this part, the alphabet A is $\{a, b, c\}$.

One can consider unavoidability in the set of square-free words (that is the set of the words which cannot be written $uzzv$ with $z \in A^+$). It can be obtained from Definition 7.4 by considering the monoid $S = \{0\} \cup T$ where T is the set of the square-free words, and the product in S is defined by

$$0.0 = 0 \quad \text{for every } t \in T, t.0 = 0.t = 0 \text{ and for every } t, t' \in T,$$

$$t.t' = \begin{cases} tt' & \text{if } tt' \text{ is square-free,} \\ 0 & \text{if } tt' \text{ is not square-free.} \end{cases}$$

Let $\mathscr{S}$ denote the set of non-ε squares (i.e. $\mathscr{S} = \{uu \mid u \in A^*\}$).

Open problems 7.6. Let X be a subset of S. Can one decide whether X is avoidable? If X is avoidable, is there a morphism ϕ from A^* to A^* such that (1) For every letter α, $|\phi(\alpha)| \geqslant 2$ (2) If u is a word which is square-free and which avoids X, then $\phi(u)$ is square-free and avoids X? Note that if such a morphism exists, then X is avoidable (unless X contains ε or a letter). Is it possible to describe unavoidable sets of a given cardinality? Minimal unavoidable sets of given cardinality seem to be finite in number, is it true?

Unavoidable language on 2 and on 3 letters

If X is an unavoidable set of words on $\{a, b, c\}$, then it can be seen that $X \cap \{a, b\}^*$ is an unavoidable language of $\{a, b\}^*$.

Open problem 7.7. Is it possible to find a minimal unavoidable language X on $\{a,b,c\}$ which does not satisfy the word-extension condition and such that $X \cap \{a,b\}^*$, $X \cap \{a,c\}^*$ and $X \cap \{b,c\}^*$, are unavoidable languages which do not satisfy the word-extension condition (respectively, on $\{a,b\}^*$, on $\{a,c\}^*$ and on $\{b,c\}^*$)?

Such an example probably exists, but should be very hard to get. Note that it has to contain at least 18 words (a power of each letter, plus at least 5 other words on $\{a,b\}^*$, 5 other ones on $\{a,c\}^*$ and 5 other ones on $\{b,c\}^*$ since a unavoidable language which does not satisfy the word-extension on 2 letters contains at least 7 words.)

References

[1] A.V. Aho, M.J. Corasick, Efficient string machines, an aid to bibliographic research, Comm. ACM 18 (6) (1975) 333–340.
[2] D.R. Bean, A. Ehrenfeucht, G.F. McNulty, Avoidable patterns in strings of symbols, Pacific J. Math 85 (2) (1979) 261–294.
[3] W. Bucher, A. Ehrenfeucht, D. Haussler, On total regulators generated by derivation relations, Proc. 12th Internat. Colloq. on Automata Languages and Programming, Lecture notes in Computer Science, vol. 194, Springer, Berlin, 1985, pp. 71–79.
[4] C. Choffrut, K. Culik, On extendibility of unavoidable sets, Discrete Appl. Math. 9 (1984) 125–137.
[5] M. Crochemore, M. Lerest, P. Wender, An optimal test on finite unavoidable sets of words, Inform. Process. Lett. 16 (1983) 179–180.
[6] L.E. Dickson, Finiteness of the odd perfect and primitive abundant numbers with r distinct prime factors, Am. J. Math. 35 (1903) 413–422.
[7] A. Ehrenfeucht, D. Haussler, G. Rozenberg, On regularity of context-free languages, Theoret. Comput. Sci. 27 (1983) 311–332.
[8] N.J. Fine, H.S. Wilf, Uniqueness theorem for periodic functions, Proc. Am. Math. Soc. 16 (1965) 109–114.
[9] G. Higman, Ordering by divisibility in abstract algebras, Proc. London Math. Soc. 2 (1952) 326–336.
[10] Kruskal, Well-quasi ordering, the tree theorem, and Vazsonyi's conjecture, Trans. Amer. Math. Soc. 95 (1960) 210–225.
[11] Kruskal, The theory of well-quasi ordering: a frequently discovered concept, J. Combin. Theory Ser. A 13 (3) (1972) 297–305.
[12] G. Kucherov, M. Rusinovitch, On ground-reducibility problem for word rewriting systems with variables, in: E. Deaton, R. Wilkerson (Eds.), Proc. 1994 ACM/SIGAPP Symp. on Applied Computing, Phoenix, ACM-Press, New York, 1994.
[13] G. Kucherov, M. Rusinovitch, Complexity of testing ground reducibility for linear word rewriting systems with variables, preprint.
[14] G. Lallement, Semigroups and Combinatorial Applications, Wiley, New York, 1979.
[15] D. Perrin, in: Lothaire (Ed.), Combinatorics on Words, Addison-Wesley, Reading, MA, 1983, Chap. 1.
[16] J. Sakarovitch, in: Lothaire (Ed.), Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.
[17] E. Ochmanski, Inevitability in concurrent systems, Inform. Process. Lett. 25 (1987) 221–225.
[18] L. Puel, Using unavoidable sets of trees to generalize Kruskal's theorem, J. Symbolic Computation 8 (4) (1989) 335–382.
[19] L. Rosaz, Unavoidable languages, cuts and innocent sets of words, Theoret. Inform. Appl. (RAIRO) 29 (5) (1995) 339–382.
[20] L. Rosaz, Unavoidable languages, Mémoire de thèse, 1993.
[21] M.P. Schutzenberger, On the synchronizing properties of certain prefix codes, Inform. and Control 7 (1964) 23–36.